THIS REPORT HAS BEEN DECLASSIFIED AND CLEARED FOR PUBLIC RELEASE.

Reprinted from JOURNAL OF EXPERIMENTAL PSYCHOLOGY,
Vol. 44, No. 5, November, 1952
Printed in U. S. A.

TWENTY QUESTIONS: EFFICIENCY IN PROBLEM SOLVING AS A FUNCTION OF SIZE OF GROUP[1]

DONALD W. TAYLOR AND WILLIAM L. FAUST

Stanford University

"Twenty Questions," popular as a parlor game in earlier years and now popular as a program on both radio and television, involves a type of problem solving that is of considerable interest psychologically.[2] To start the game, the participants are told only whether the object they are to attempt to identify is animal, vegetable, or mineral. In searching for the object which is the solution to the problem, they ask a series of questions, each of which can be answered "Yes" or "No." To find the solution most economically, they must use a high order of conceptualization, gradually increasing the specificity of the concepts employed until they arrive at the particular object.

The game is of psychological interest first of all because it appears to involve a type of problem solving more similar to much problem solving in everyday life than that ordinarily studied in psychological experiments. The solution is obtained not by a series of rigorous well-defined steps. Rather one starts with a general, somewhat vague problem. Questions are asked and information obtained. Upon the basis of this information, new questions are formulated. This procedure continues until the problem is solved. This type of problem solving is also of interest because it seems more similar to much of the problem solving in scientific research than does that involved in problems susceptible of rigorous, deductive mathematical or logical solution.

The use of the game in psychological experiments is recommended by several other considerations: It is quite interesting to college undergraduates; motivation is easily sustained for a period of several days. A very large number of problems of this kind are available. The same problems can be used with children and with adults. The same problems are appropriate for use with individuals and with groups of varying size.

The present experiment, the first in a series planned using the game, was designed to answer three questions: (*a*) How rapidly is the skill involved in the game learned? (*b*) How does efficiency in solving this type of problem vary as a function of the size of the group participating? (*c*) Does improvement in individual performance occur more rapidly with individual practice or with practice as a member of a group?

The second of these three questions is perhaps the most interesting. For many kinds of work, it seems quite reasonable that if a particular job must be completed in a shorter time, the number of people in the group working on it should be increased. It

[1] This experiment was carried out under Project NR 192-018 supported by Contract N6 onr-25125 between the Office of Naval Research and Stanford University. The first author designed the experiment, supervised the analysis of the data, and prepared the present report. The second author conducted the experiment and carried out the analysis of the data. Work on the contract is under the direction of the first author.

[2] The idea of using "Twenty Questions" in experimental studies of problem solving is not new. As was discovered after the present study was partly completed, Lindley (3) suggested the use of the game for this purpose in an article published in 1897.

is not clear that increasing the size of a group engaged in solving a problem will necessarily reduce the time required for its solution. Indeed, it appears likely that in some cases it will actually increase the time required. Shaw (4) has presented data which indicate that the performance of groups of four is superior to that of individuals. However, further experimentation with larger samples, varying size groups, and different types of problems is needed to determine adequately the relation between group size and efficiency in problem solving.

PROCEDURE

A total of 105 students from the elementary course in psychology served as *S*s. The *S*s were assigned by chance to work in solving the problems either alone, in pairs, or as a member of a group of four. There were 15 individual *S*s, 15 groups of two, and 15 groups of four. Each individual or group was given four problems a day for four successive days. On the fifth day, all *S*s worked alone, each being given four problems.

From a longer list of objects originally constructed, 60 were selected for use as problem topics. Included were 20 animal, 20 vegetable, and 20 mineral objects. Excluded were objects which did not clearly fit in only one of the three categories; e.g., hammer was not included because, with a handle of wood and a head of metal, it would be classed as both vegetable and mineral. Also excluded were objects which could not be expected to be familiar to almost every college student. Examples of objects included are: newspaper, Bob Hope, scissors, camel, dime, rubber band.

With four problems a day for five days, a total of only 20 problems was needed for presentation to any particular *S* or group. However, to minimize the possibility that an *S* would have any knowledge of what problem object to expect, it was decided to use a total of 60 different objects. This precaution seemed desirable although the instructions to be given all *S*s specifically requested that they not discuss the problems with other students. It should be added that no evidence was obtained during the course of the experiment to indicate that any *S* had previously heard mentioned a problem object he was to be given.

Since the nature of the learning curve was of interest, it was necessary to control the order of presentation of the problems in such a way that those given on any one day would be equal in difficulty to those given on any other day. In the absence of any measure of the difficulty of the individual problems, the following procedure was employed: The 20 animal objects were listed in chance order, as were the 20 vegetable and the 20 mineral objects. To obtain a group of four for use the first day, the first item was taken from each of the three lists together with the next item from one of the three chosen by chance. Similarly, to obtain four objects for use the second day, the next item was taken from each of the three lists; the fourth item was then obtained by taking the next in order on one of the two lists from which the extra item had not been taken the first day. This procedure was repeated to provide four problems for the third, fourth, and fifth days. A second and a third set of four problems for each of five days were obtained by continuing the same procedure. Next the three lists of 20 were individually reshuffled and the entire procedure repeated to obtain a fourth, fifth, and sixth set.

In the experiment, the first, seventh, and thirteenth individual, pair, or group of four *S*s received the first set of problems. The second, eighth, and fourteenth received the second set, and so on. As a result of this procedure, the order and the frequency of appearance of the problems were the same for individual *S*s as for groups of two or of four.

All *S*s were told that both the number of questions and the time required to reach solution would be recorded, but it was emphasized that number of questions was the more important score. In presenting each problem, *E* stated simply whether the object sought was animal, vegetable, or mineral. Time was measured by means of a stopwatch. A special data sheet was used for groups of two and of four to record which *S* asked each question. To each question, *E* replied "Yes," "No," "Partly," "Sometimes," or "Not in the usual sense of the word." If the question could not be answered in one of these ways or was unclear, *S* was asked to restate it.

The instructions given to groups of two or of four made clear that they might talk freely to each other, reviewing answers to previous questions or suggesting possible questions to ask. It was emphasized that they were not to compete against each other, but were to cooperate as a group to get the answer; they were told that the efficiency of their group would be compared with that of other groups.

As the name of the game indicates, *S*s are traditionally allowed 20 questions in which to obtain the solution. Pretesting showed, however, that with naive *S*s this limit results in a rather large proportion of failures. Accordingly,

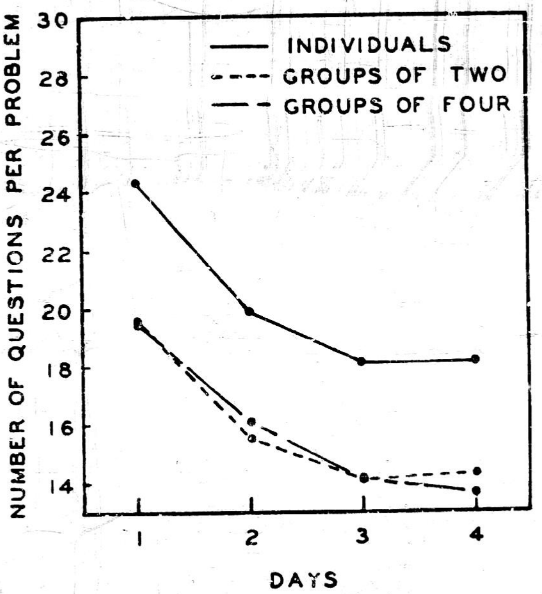

FIG. 1. Number of questions per problem as a function of days of practice and of size of group

to simplify the analysis of the data to be obtained, the number of questions permitted was increased to 30. Examination of the distributions of scores obtained suggests that, at least after the first day, the performance of individuals or groups of *S*s who do not reach solution in 30 questions is qualitatively different from that of those who do. The *E*'s impression is that in most cases of failure there was established an incorrect set which was unchanging even in the face of answers irreconcilable with it; it seemed that in such cases the *S*s might easily have asked 50 or 60 questions without solving the problem.

Results

Rate of learning.—The first question the experiment was designed to answer concerned the speed of learning of the skill involved. The data in Fig. 1 show that there is rapid improvement in the performance of both individuals and groups. By the fourth day the curves appear already to be flattening out. The score for an individual or single group for one day was the median of the number of questions required to solve each of the four problems on that day. The median was used instead of the mean because there were some failures. Each point plotted in Fig. 1 is the mean of these median scores on one day for 15 individuals, or for 15 groups of two or of four. In those few cases where an individual or group failed two or more problems on a single day, the median was obtained by treating the failures as though solution had been reached in 31 questions; the number of such cases was too small to affect the results appreciably; after the first day there were no such cases except among individual *S*s and even there they were rare.

The mean number of failures per problem on each day by individuals or groups is shown in Fig. 2. Thus, for example, on the first day the mean number of failures per problem among the 15 groups of four was .08; in other words, about one-twelfth of the problems were failed. The improvement in performance over four days in terms of number of failures per problem is consistent with that shown in Fig. 1 in terms of number of questions per problem solved.

Figure 3 shows the decrease over four days in the amount of time

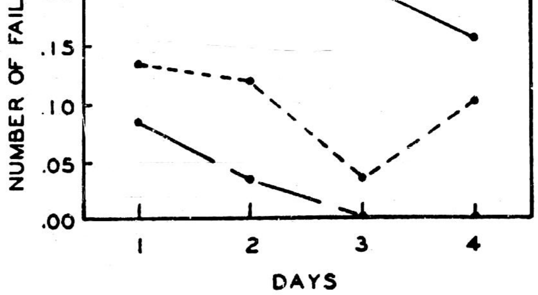

FIG. 2. Number of failures per problem as a function of days of practice and of size of group

required per problem. The time required, of course, is somewhat dependent on the number of questions asked, although not entirely so. The score for an individual or single group for one day was the median time required for solution of the four problems. In those few cases where there were two or more failures in one day, the median of the four times was taken simply as obtained; this procedure underestimates somewhat the median time that would have been required to solve all four problems, but as before the number of such cases was too small to affect the general results appreciably.

Size of group.—The second and major question with which the experiment was concerned involved the relation between efficiency in problem solving and size of group. As is evident in Fig. 1, there was no significant difference between groups of two and groups of four in terms of the number of questions required to reach solution. The performance of individuals working alone, however, was consistently inferior to that of either size group. The t technique was used to test the difference on each day between the mean score of the 15 individuals and the mean score of the 15 pairs of Ss, and also that of the 15 groups of four. The values of t obtained are presented in Table 1. With 28 df, a t

TABLE 1

VALUES OF t FOR DIFFERENCES BETWEEN MEAN SCORES: NUMBER OF QUESTIONS PER PROBLEM

Day	Individuals versus Groups of Two	Individuals versus Groups of Four
1	2.67	2.18
2	2.86	1.96
3	2.30	2.22
4	2.11	2.45
All 4	2.64	2.62

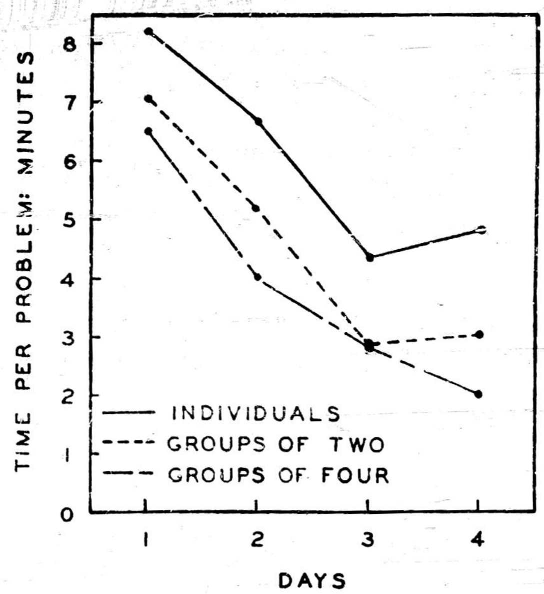

FIG. 3. Time per problem as a function of days of practice and of size of group

of 2.05 is required for significance at the .05 level and of 2.76 at the .01 level. All of the differences but one are significant at or beyond the .05 level.

A score for all four days was obtained for each individual or single group by taking the median number of questions required to solve the 16 problems. In terms of the means of these scores, the performance both of groups of two and of four is significantly better (.02 level) than that of individuals working alone (see Table 1).

That there were differences as a function of group size in terms of number of failures to reach solution is suggested by Fig. 2. Because of the fact that, as would be expected, the distributions of failure scores were not normal, t could not be used to test the significance of these differences. Instead a test described by Festinger (2) was employed. The mean number of failures per problem, all four days included, was for individuals, .26; for pairs, .10; for groups of four, .03. The values of d obtained indicate that

TABLE 2

VALUES OF t FOR DIFFERENCES BETWEEN MEAN SCORES: TIME PER PROBLEM

Day	Individuals versus Groups of Two	Individuals versus Groups of Four	Groups of Two versus Groups of Four
1	.85	1.14	.12
2	1.01	2.36	.93
3	2.20	2.22	.06
4	2.15	3.49	1.90
All 4	2.39	3.27	1.18

the difference between individuals and groups of four is significant at well beyond the .01 level; the difference between individuals and pairs and the difference between pairs and groups of four are both significant at about the .02 level.

Differences in mean time to solution among individuals, groups of two, and groups of four may be seen in Fig. 3. Fortunately, the distributions of the median times, of which the individual points plotted in Fig. 3 are the means, were such as to make the use of t appropriate in testing the significance of differences between means. Table 2 presents the values of t obtained for the various comparisons. As in the case of number of questions required, none of the differences between groups of two and of four is significant. Differences between individuals and groups of two on the third and fourth days are significant at the .05 level; differences between individuals and groups of four on all except the first day are significant at the same level or beyond.

A score for all four days was obtained for each individual or single group by taking the median time required for the 16 problems. The means of these scores were 5.06 for individuals, 3.70 for groups of two, and 3.15 for groups of four. The values of t given in Table 2 show that the difference between the first and second mean is significant at the .05 level, and between the first and third mean at the .01 level.

Group performance was superior to individual performance in terms of elapsed time to solution. However, if, instead, an analysis is made in terms of number of man-minutes required for solution, the nature of the results obtained changes sharply. The number of man-minutes for a problem will, of course, be equal to the elapsed time multiplied by the number of persons in the group. In terms of man-minutes, the mean of the scores for all four days was 5.06 for individuals, 7.40 for groups of two, and 12.60 for groups of four. Since the variances for these three means were clearly not homogeneous, the use of t was not appropriate for testing the significance of the obtained differences. Instead, t' was employed (1). Both the difference between individuals and groups of two and the difference between groups of two and groups of four are significant at the .02 level. The difference between individuals and groups of four is significant at the .001 level. Clearly, in terms of man-minutes, the performance of individuals was superior to that of groups of two or of four; in addition, the performance of groups of two was superior to that of groups of four.

A supplementary question of some interest is whether the member of a group of two or of four getting the correct answer asked significantly more questions than the other member or members of the group. An analysis for all four days combined showed that for groups of two, the individual getting the correct answer asked an average of 1.55 questions more than the individual who failed to get the answer. A t of 5.04 with 14 df shows this to be significantly different from zero at the .001 level. However, it may be plausibly argued that in making this comparison, the final question which identified the correct object should be excluded. Before asking it, the individual had correctly formulated the answer. If the final answer is excluded, the difference is reduced from 1.55 to .55. This yields a t of 1.74 and is not significantly different from zero.

A similar analysis was done for groups of four. When the final question is included, the mean difference between the number of questions asked by the individual getting the answer and the average number asked by the other three members was 1.53. With a t of 6.50, this is significantly different from zero at the .001 level. Excluding the final question reduces the mean difference to .53. However, with a t of 2.25, this is still significantly different from zero at the .05 level. There appears to be some tendency for the member of a group of four getting the correct answer to ask more questions, even excluding the final question, than do other members of the group.

Individual versus group practice.—The third question which the experiment was intended to answer was whether improvement in individual performance occurs more rapidly with individual practice or with practice as a member of a group. To answer this question, all *Ss* worked alone on the fifth day. As before, the score for each individual was the median number of questions required to solve the four problems. The mean of these scores for the 15 *Ss* who had previously worked alone was 20.8;[3] for the 30 who had worked in pairs, 19.3; and for the 60 who had been members of groups of four, 19.1. None of the differences among these means is significant. Nor were any of the differences significant among the corresponding means on the fifth day for number of failures or for time scores. Learning went on as well in groups of two or of four as in individual practice.

Discussion

The results obtained show that there is rapid learning of the skill involved in the game. The question now arises as to just what it is that is learned. To determine this, a qualitative analysis of the kinds of questions asked on successive days will be necessary. In a second experiment, now in progress, a complete record of all questions asked is being made in order that such an analysis can be carried out.

Group performances were superior to individual performance in terms of number of questions, number of failures, and elapsed time per problem; but the performance of groups of four was not superior to that of groups of two, except in terms of the number of failures to reach solution. Whether one could confidently have predicted such group superiority is questionable: Individual members of the group might have failed to make effective use of the information yielded by questions asked by other members; if this had been the case, the number of questions required by a group would have been larger, rather than smaller, than that required by an individual.

The fact that there were negligible differences between groups of two and of four either in number of questions

[3] Comparison of this mean for the fifth day with that for the fourth day (20.8 versus 18.1) shown in Fig. 1 may raise the question: Why should the performance on the fifth day be inferior to that on the fourth day in view of the fact that the conditions under which these 15 individuals worked were the same on both days? However, the difference between these two means is not significant (t = 1.04).

or in elapsed time strongly suggests that the optimum size group is not larger than four. Proof of this will require further experimentation with other size groups. Additional experiments are also needed to determine whether the optimum size group is similar for other types of problems.

The question may be raised as to why there was a significant difference between groups of two and of four in number of failures to reach solution, this in spite of the fact that there were negligible differences in number of questions or elapsed time. A possible explanation is that increasing the number of participants from two to four reduces the probability of a persisting wrong set resulting in complete failure. For an individual, a wrong set once established may make it impossible to solve the problem. The probability that a wrong set would be established simultaneously for all participants would be smaller for a group of four than for a group of two.

Although group performances were superior to individual performance in terms of elapsed time to solution, the performance of individuals was superior to that of either size group in terms of number of man-minutes required for solution. The practical implications of this fact should not be overlooked. It appears probable that there are many kinds of problems which a group will solve more quickly than an individual. If elapsed time in hours, weeks, or months is the primary consideration, then such problems should be undertaken by groups. However, it appears equally probable that few of those same problems will be solved more efficiently in terms of man-minutes or man-hours by groups than by individuals. If a group of two is to solve a problem more efficiently than an individual in these latter terms, it must solve it in less than half the elapsed time required by the individual. Similarly, a group of four to be more efficient must solve the problem in less than one-fourth the elapsed time required by the individual. The importance of this point appears to be frequently overlooked.

What it is that accounts for the superiority of group as compared to individual performance in terms of number of questions or elapsed time remains to be determined. The suggestion may be made that the superiority of the group is due to the performance of the best member of the group. If one were to pick the most able individual from each of 15 groups of four, it would be expected that the performance of these 15 individuals would be superior to that of 15 individuals chosen by random sampling. The mean number of questions required by groups of four on the fourth day was 13.6. The mean of the best individual performances on the fifth day by former members of each of the 15 groups of four was 14.8, not significantly different from 13.6. This fact would seem to support the suggestion just made. However, this comparison is not fully valid. Which former member of a group of four had the best performance on the fifth day very probably depended partly on ability and to a considerable extent on chance. Selecting the best individual performance from each of the 15 groups thus capitalizes on chance in a way that reduces the mean obtained; it may yet be true that the mean performance of the 15 groups would be superior to that of the best individuals in each of the 15 groups.

That the superior performance of the group is not simply a function of the performance of the best member of the group is suggested by another consideration. If this were the case, then the larger the group, the better on the average should be the performance of the best member on the basis of sampling alone; hence the larger the group, the better should be the performance. The negligible differences obtained between groups of two and of four fail to confirm this expectation.

It may be expected that other factors such as broader range of relevant information, greater flexibility in approach, etc., are at least partly responsible for the superiority of group over individual performance. What these factors are and how they operate to produce an optimum size for a group can be determined only by additional experimentation.

An interesting supplementary question is whether the member of a group who obtains the right answer does so largely because he asks more questions than the other members of the group. The data obtained show that the number of questions asked by the member of a group of two obtaining the correct answer does not differ significantly from the number asked by the other member. A difference significant at the .05 level was found between the number asked by the member of a group of four obtaining the correct answer and the mean number asked by the other three members. However, this significant difference was only a matter of .53 questions per problem. It seems doubtful that getting the right answer is primarily due to the asking of more questions either in groups of two or of four.

The results obtained on the fifth day showed that learning resulting in improvement in individual performance occurred as rapidly with individual practice as with practice as a member of a group of two or of four. This fact, of course, should not be taken to mean that improvement is qualitatively the same under the different conditions. It may or may not be.

Summary and Conclusions

The game of "Twenty Questions" was employed in an experiment on problem solving. A total of 105 *Ss* were assigned by chance to solve such problems working either alone, in pairs, or in groups of four. There were 15 individual *Ss*, 15 groups of two, and 15 groups of four. Each individual or group was given four problems a day for four successive days. On the fifth day, all *Ss* worked alone, each being given four problems.

Both the number of questions and the time required to solve each problem were recorded. Problems not solved in 30 questions were counted as failures.

1. In terms of number of questions, rapid improvement occurred in the performance both of individuals and of groups. By the fourth day, the curves appeared to be flattening out. Similar results were obtained in terms both of number of failures and of time per problem.

2. Group performances were superior to individual performance in terms of number of questions, number of failures, and elapsed time per problem; but the performance of groups of four was not superior to that of groups of two, except in terms of the number of failures to reach solution.

3. In terms of man-minutes required for solution, the performance of individuals was superior to that of groups; the performance of groups of two was superior to that of groups of four.

4. Improvement in individual performance occurred as rapidly with individual practice as with practice as a member of a group.

(Received for priority publication August 18, 1952)

References

1. Cochran, W. G., & Cox, G. M. *Experimental designs.* New York: Wiley, 1950.
2. Festinger, L. The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 1946, **11**, 97–105.
3. Lindley, E. H. A study of puzzles with special reference to the psychology of mental adaptation. *Amer. J. Psychol.*, 1897, **8**, 431–493.
4. Shaw, M. E. A comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, **54**, 491–504.